Streaming Software Analytics

Georgios Gousios
Radboud University Nijmegen
Nijmegen, the Netherlands

Dominik Safaric
Radboud University Nijmegen
Nijmegen, the Netherlands

Joost Visser
Radboud University Nijmegen
Nijmegen, the Netherlands
A service at Microsoft generates about 12 billion lines of log messages each day for indecent management.

Kafka handles over 55 billion message writes per day with averaging 578,000 messages per second.

Nearly all sectors in the US economy had at least an average of 200 terabytes of stored data per company.

Data-driven decision making.
Software analytics aim at utilising data-driven approaches to enable software practitioners to perform data exploration and analysis in order to obtain insightful and actionable information.

Dongmei Z, Shi H. Yingnong D., Software analytics in practice, IEEE, 2013
Limitations?
Integration of data sources.
Statically defined data aggregations.
Conformance to **multiple stakeholder group** information needs.
Streaming software analytics.
An infrastructure supporting for an unified data aggregation and summarisation query method derived from data streams of historical and real-time analytical data.
Data sources

Push & pull based data sources

Streaming software analytics

Data silo

Streaming and querying engine

Aggregation and summarisation
Streaming and querying engine

Data silo → Streaming and querying engine → Aggregation and summarisation

Push & pull based data sources
Streaming and querying engine

Data silo

Push & pull based data sources

Streaming software analytics

Aggregation and summarisation
Streaming and querying engine

Data silo

Streaming and querying engine

Aggregation and summarisation

Push & pull based data sources

Data sources

Streaming software analytics
Streaming and querying engine

Push & pull based data sources

Data silo

Streaming and querying engine

Aggregation and summarisation

Data sources

Streaming software analytics
Prototype
Challenges (so far) encountered

> Selection of a stream processing engine
Challenges (so far) encountered

> Selection of a stream processing engine
> Extensibility of the infrastructure
Challenges (so far) encountered

> Selection of a stream processing engine
> Extensibility of the infrastructure
> Initial data source selection
Integrate

Query

Aggregate

@dominiksafaric[at]gmail.com